

An analysis of the MGP edge data

David Joyner*

2010-08-11

In June 2010, Mitch Keller of the MGP kindly sent the author “edge data” from the Mathematics Genealogy Project as a csv file. This was a list of 144892 pairs, the first entry being the advisor (or his/her number as designated by the MGP) and the second the advisee. No real names were included. This note contains some analysis of this data.

The MGP website has the following statement: “143840 records as of 2 August 2010,” so the edge set seems to contain a small number of duplicates but be slightly larger than the data in the website. Reading in the provided csv file and noting that the first and last line are extraneous, we see that there are indeed 144892 advisor-advisee pairs in the list.

```
sage: edge_list = open("../research/mgp_edges.csv", "r")
sage: edges_all = edge_list.readlines()
sage: n = len(edges_all); n
144894
```

Now, regard each pair as an edge in a graph where two vertices are connected by an edge if and only if one is the advisor of the other. There are 33206 advisor vertices V_1 and 126556 advisee vertices V_2 . These are overlapping sets whose union $V_1 \cup V_2$ has 135498 vertices, so there are $135498 - 126556 = 8942$ advisors who are not advisees of someone else in the database.

*The author thanks the Mathematics Genealogy Project (<http://genealogy.math.ndsu.nodak.edu/>) for providing data from its database for use in this research. Contact info: Math Dept, USNA, Annapolis MD 21402, wj@usna.edu

There are 3714 connected components in this graph. The largest connected component G_0 has 121424 vertices and the smallest has 2 vertices. There are 1937 connected subgraphs having exactly 2 vertices. The average degree of G_0 is about 2.21 and girth 3.

How many advisors who are not advisees of someone else in the database are there in the component G_0 ? In G_0 , there are 27849 advisor vertices V_{01} and 116785 advisee vertices V_{02} . These are overlapping sets whose union $V_{01} \cup V_{02}$ has ??? vertices, so there are $121424 - 116785 = 4639$ advisors in G_0 who are not advisees of someone else in the database.

What are the sizes of the other components, with frequency of size? Even though there are well over 3000 connected components of G , there are only about 45 different sizes represented among those subgraphs. Here is a list of pairs (x, y) , where x represents the size of a connected subgraph and y counts the number of connected subgraphs of that size.

[(2, 1937), (3, 761), (4, 327), (5, 188), (6, 104), (7, 69), (8, 70), (9, 47), (10, 27), (11, 40), (12, 22), (13, 19), (14, 14), (15, 9), (16, 7), (17, 12), (18, 7), (19, 6), (20, 3), (21, 5), (22, 4), (23, 3), (24, 2), (25, 4), (26, 3), (27, 4), (29, 3), (30, 2), (31, 1), (32, 1), (33, 1), (34, 2), (35, 1), (36, 1), (38, 1), (40, 1), (42, 1), (45, 1), (61, 1), (79, 1), (128, 1), (121424, 1)].

The list plot of these points, minus the four points with with two largest x -values and the two largest y -values, is given in Figure 1.

The degree centrality of G_0 is the list parameterized by the vertex set V_0 of G_0 whose v -th entry is the fraction of vertices connected to $v \in V_0$. The centrality of a vertex within a graph determines the relative importance of that vertex to its graph. Degree centrality measures the number of edges incident upon a vertex.

The maximum degree centrality for G_0 is 0.0008729..., associated to vertex 93643. Using the “quick search” feature on the site <http://www.genealogy.ams.org/>, this vertex denotes Professor C.-C. Jay Kuo of the University of Southern California, who has 104 advisees.

The closeness centrality is defined to be

$$\frac{1}{\text{average distance to all vertices}}.$$

Closeness centrality is an inverse measure of centrality in that a larger value indicates a less central vertex while a smaller value indicates a more central

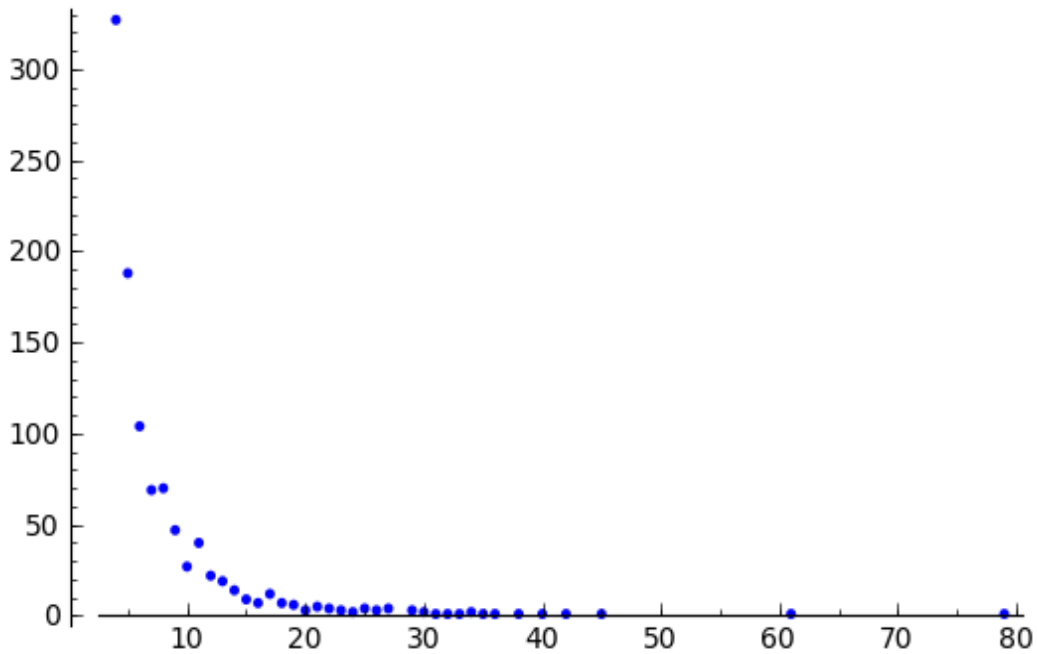


Figure 1: The sizes of the connected components of the graph, minus two highs and 2 lows.

vertex. However “Professor 93643” (=Professor Kuo) has closeness centrality of 0.09014..., which is not the maximum nor minimum value (although the extrema are not known for G_0). The maximum closeness centrality of G_0 is at least 0.11 and the minimum is no more than 0.05. Computing the closeness centrality for the first 500 vertices took about 16 hours on a mac pro.

References

- [S] W. Stein, **Sage - a mathematical software package**, version 4.5.1, <http://www.sagemath.org/>.